CS657 Fall 2017

FINAL EXAM

NAME: Robert Jarvis

G#: G00647058

1. Given the following characteristic matrix for a set of 4 documents and 5 shingles,

Shingle	S1	S2	S3	S4
0	1	0	1	0
1	1	1	0	0
2	0	1	0	1
3	1	0	1	0
4	0	0	0	1

Answer the following questions

   a. Which pair of documents are the most similar (using Jaccard similarity)?
   b. Which documents contain shingle 0?
   c. Using the hashing functions:
      $h_1(n) = 3n + 2 \ mod \ 7, h_2(n) = n - 1 \ mod \ 7$, compute the matrix of signatures for the documents above. To what permutations do these hash functions correspond?
   d. Using the matrix of signatures and the algorithm covered in class, compute the signature matrix and compare the similarities found by using that matrix with the original similarities for the documents.

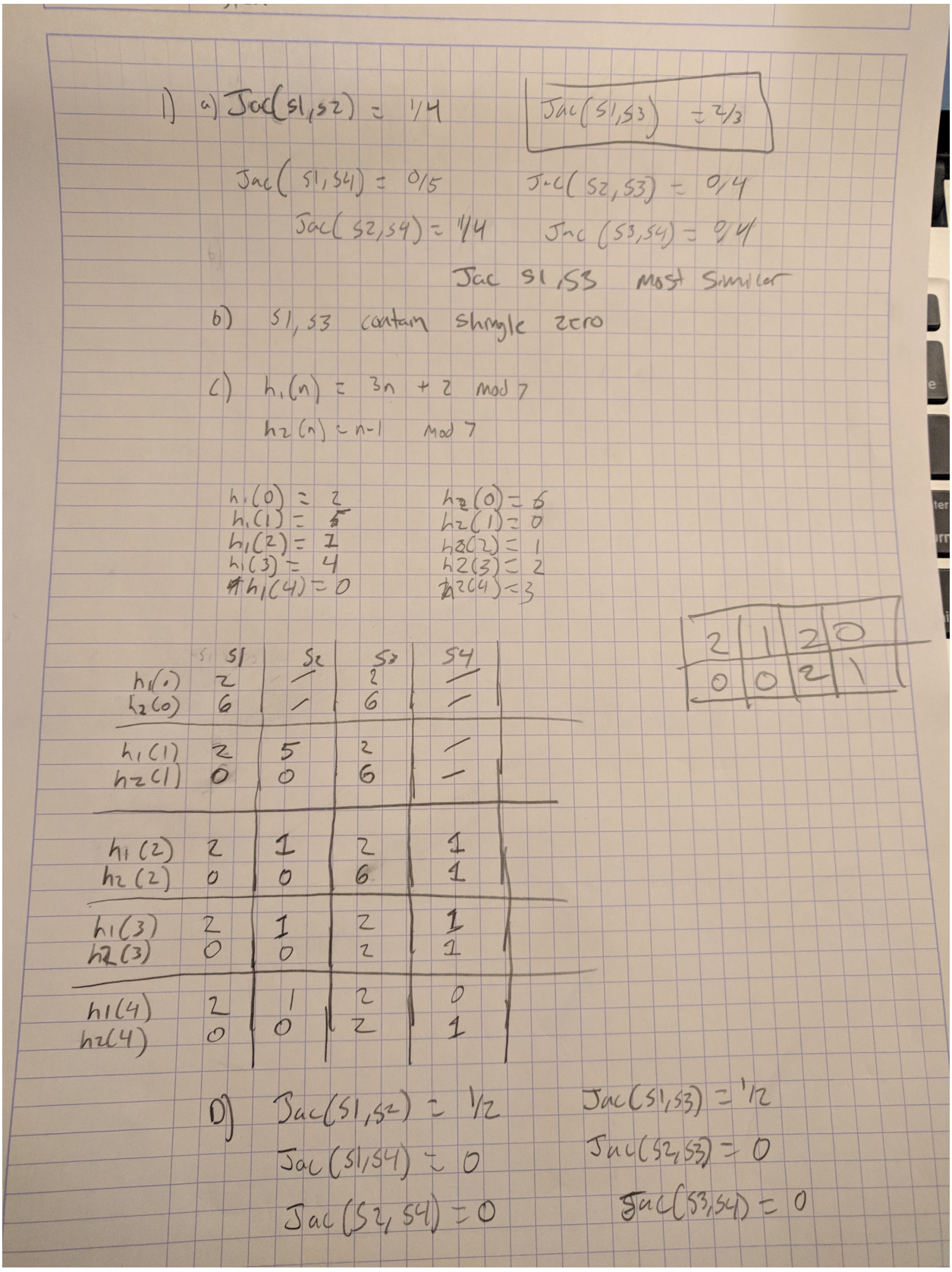


Figure 1 Question One

2. Starting with a $(0.1,0.8,0.9,0.1)$ family, what are the results of amplifying it using the following constructs?
    a. A 4-way AND followed by a 5-way OR
    b. A 5-way OR followed by a 4-way AND
    c. Which of the two has a large False Positive rate?

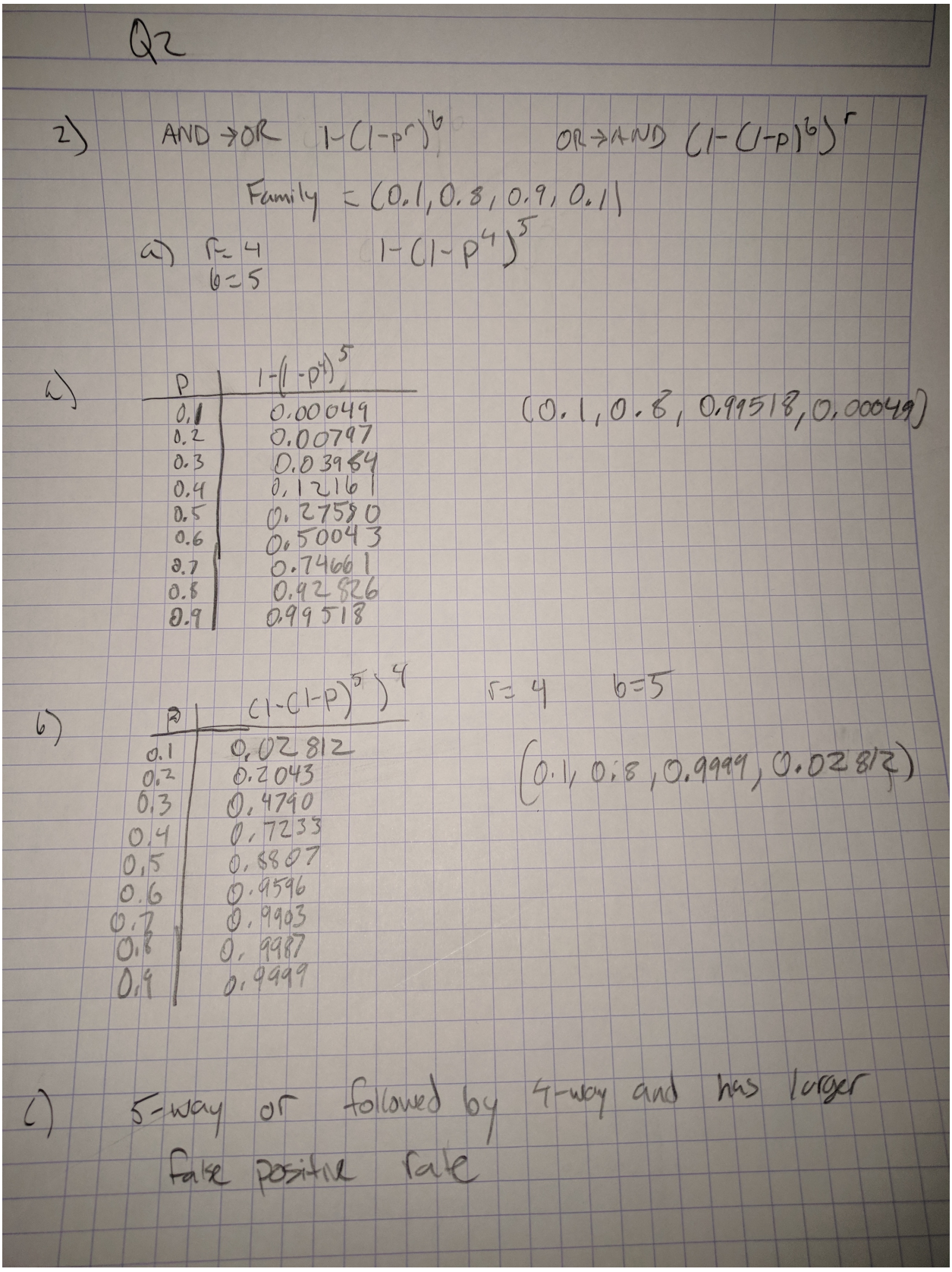


Figure 2 Question Two

3. Compute sketches of the vectors [2,3,4,5],[-2,3,-4,5] and [2,-3,4,-5} using the following random vectors $v_1 = [1,1,1,-1]$, $v_2 = [1,1,-1,-1]$, $v_3 = [1,1,1,1]$, $v_4 = [-1,1,1,1]$. Compute the true and estimated cosines between the vectors .

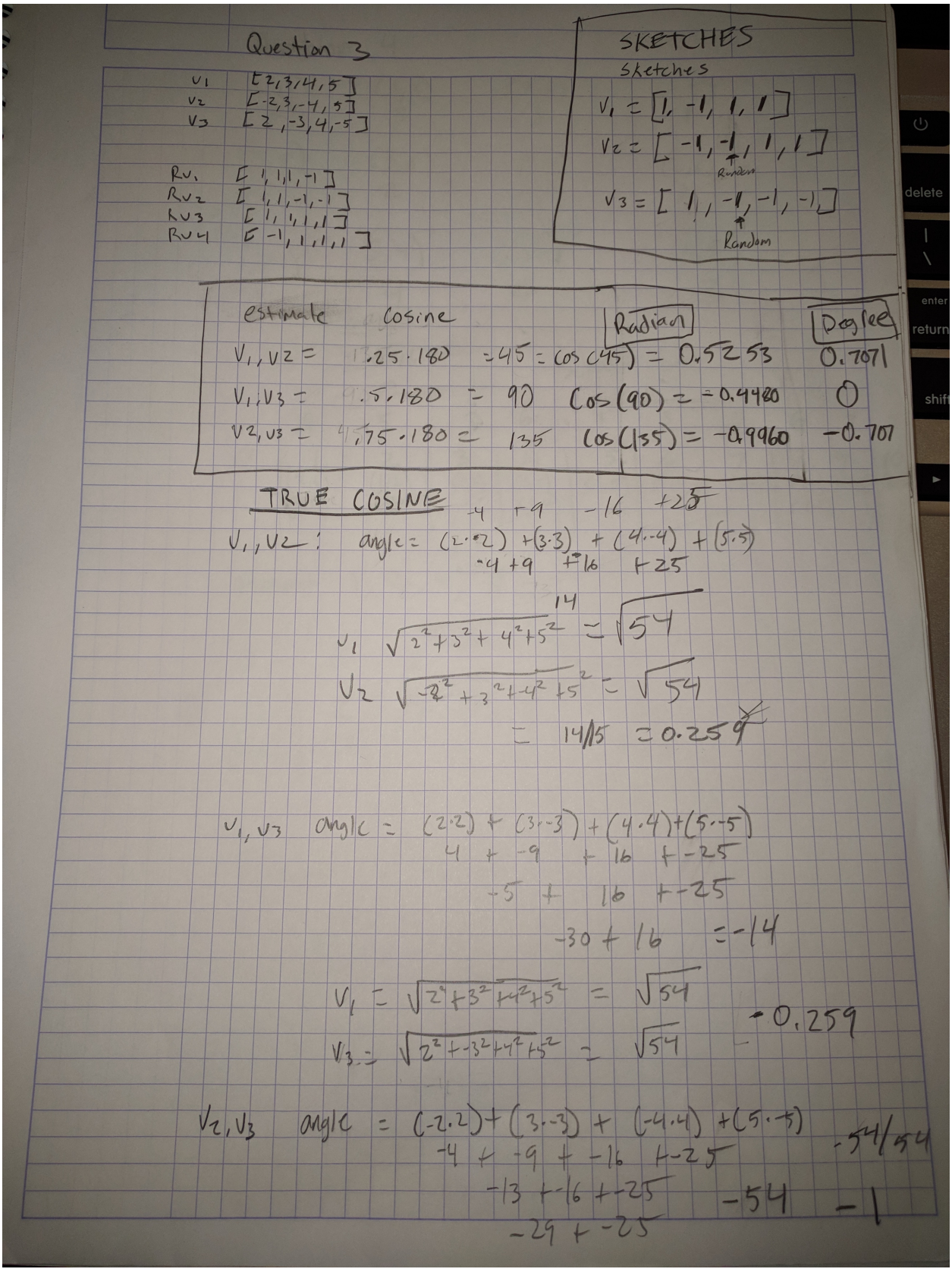


Figure 3 Question Three

4. Consider the following collection of baskets

   {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}
   {1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5}
   {3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}

   And a support threshold of 4. We are using PCY and a hash table of 11 buckets where the pair (i,j) is hashed to ixj mod 11

   In the second stage, we hash pairs to 9 buckets using i+j mod 9.

   Answer the following questions:

   a. Which buckets in the first pass are frequent?
   b. Which pairs are counted on the second pass?
   c. What are the counts of the buckets on the second pass?
   d. Does the second pass reduce the number of candidates?

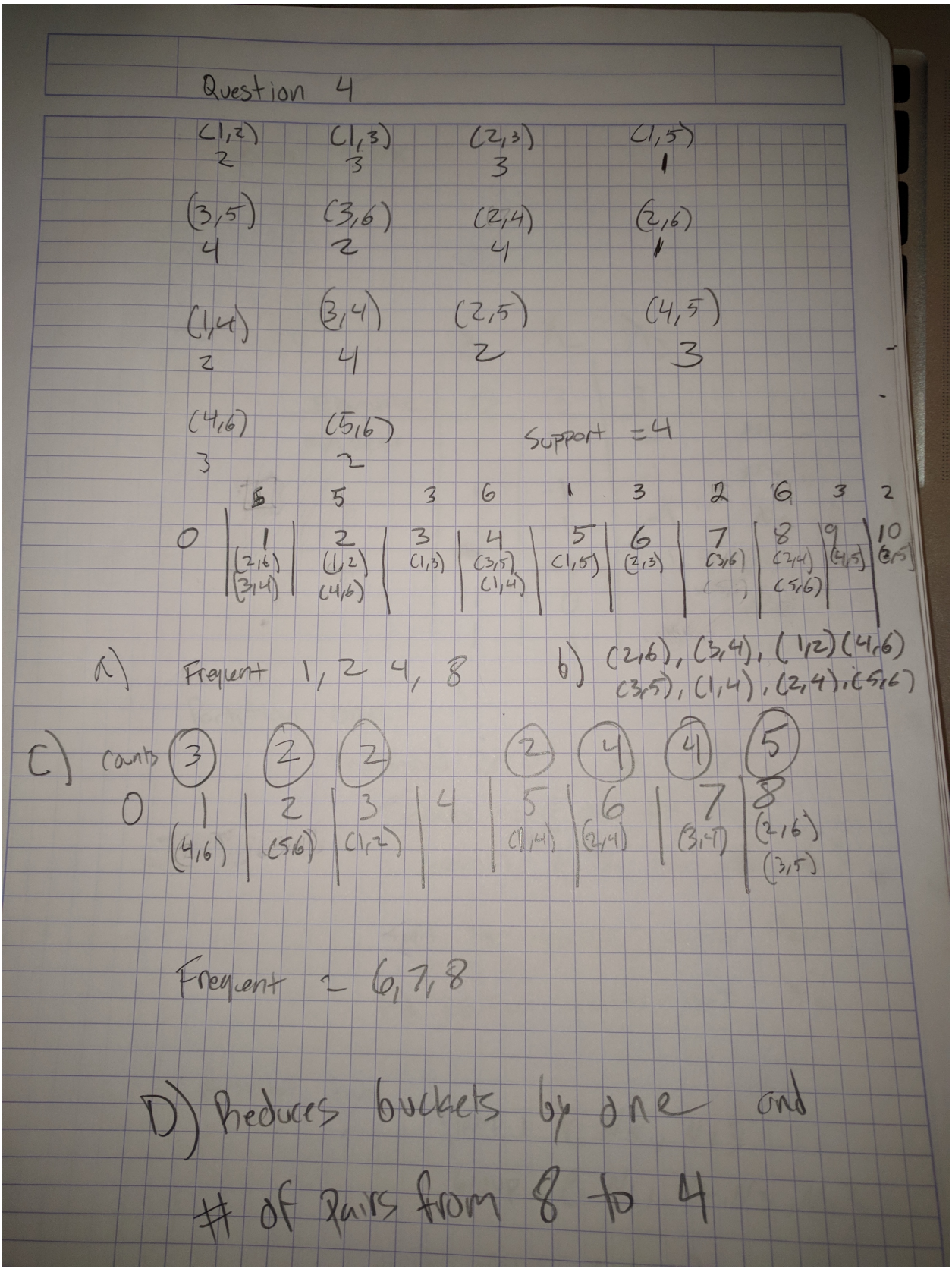
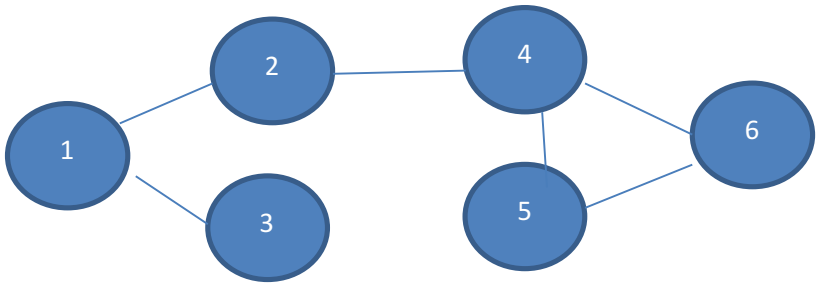


Figure 4 Question Four

5. Consider the graph



Compute the Laplacian matrix, find its first two eigenvectors and suggest two different partitions of the graph based on those eigenvectors.

Question 5

Laplacian Matrix

	1	2	3	4	5	6
1	2	-1	-1	0	0	0
2	-1	2	0	-1	0	0
3	-1	0	1	0	0	0
4	0	-1	0	3	-1	-1
5	0	0	0	-1	2	-1
6	0	0	0	-1	-1	2

$$\begin{vmatrix} 2-\lambda & -1 & -1 & 0 & 0 & 0 \\ -1 & 2-\lambda & 0 & -1 & 0 & 0 \\ -1 & 0 & 1-\lambda & 0 & 0 & 0 \\ 0 & -1 & 0 & 3-\lambda & -1 & -1 \\ 0 & 0 & 0 & -1 & 2-\lambda & -1 \\ 0 & 0 & 0 & 0 & -1 & 2-\lambda \end{vmatrix}$$

$\lambda_1 \quad \lambda_2 \quad \lambda_3$

LAPLACIAN MATRIX

Eigenvalue	0	3	3	
Eigen vector	1	2	2	N_1
	1	-1	-1	N_2
	1	-1	-1	N_3
	1	-1	-1	N_4
	1	0	1	N_5
	1	1	0	N_6

Partition one

threshold zero on λ_2

Group 1 = { N_1, N_5, N_6 }

Group 2 = { N_2, N_3, N_4, N_3 }

split the negative values from the positive values on λ_2

Partition two

group 1 { N_5, N_6 }

group 2 { N_1, N_2, N_3, N_4 }

split values across λ_1 and λ_2

Keep 0's and 1's together